Assignment 3: Data Exploration

Yared S. Asfaw, Section #03

OVERVIEW

This exercise accompanies the lessons in Environmental Data Analytics on Data Exploration.

Directions

1. Change "Student Name, Section #" on line 3 (above) with your name and section number.
2. Work through the steps, **creating code and output** that fulfill each instruction.
3. Be sure to **answer the questions** in this assignment document.
4. When you have completed the assignment, **Knit** the text and code into a single PDF file.
5. After Knitting, submit the completed exercise (PDF file) to the dropbox in Sakai. Add your last name into the file name (e.g., "FirstLast_A03_DataExploration.Rmd") prior to submission.

The completed exercise is due on <>.

Set up your R session

1. Check your working directory, load necessary packages (tidyverse), and upload two datasets: the ECOTOX neonicotinoid dataset (ECOTOX_Neonicotinoids_Insects_raw.csv) and the Niwot Ridge NEON dataset for litter and woody debris (NEON_NIWO_Litter_massdata_2018-08_raw.csv). Name these datasets "Neonics" and "Litter", respectively. **Be sure to add the `stringsAsFactors = TRUE` parameter to the function when reading in the CSV files.**

```
getwd()
```

```
## [1] "E:/EDA/Environmental_Data_Analytics_2022/Assignments"
```

```
# Load packages
library(tidyverse)

# Importing the data set
neonics.data <- read.csv(file = "../Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv",
header = TRUE,sep = ",", stringsAsFactors = TRUE)
```

```
is.data.frame(neonics.data)
```

```
## [1] TRUE
```

```
litter.data <- read.csv(file = "../Data/Raw/NEON_NIWO_Litter_massdata_2018-08_raw.csv",
header = TRUE,sep = ",", stringsAsFactors = TRUE)
```

```
is.data.frame(litter.data)
```

```
## [1] TRUE
```

Learn about your system

2. The neonicotinoid dataset was collected from the Environmental Protection Agency's ECOTOX Knowledgebase, a database for ecotoxicology research. Neonicotinoids are a class of insecticides used widely in agriculture. The dataset that has been pulled includes all studies published on insects. Why might we be interested in the ecotoxicologoy of neonicotinoids on insects? Feel free to do a brief internet search if you feel you need more background information.

   Answer: Neonicotinoids are now the most widely used insecticides in the world and can be used for many different crop types. They can be sprayed onto foliage or applied as soil drenches, but they are predominantly used as seed treatments. When used this way, neonicotinoids are taken up by all parts of the plant as it grows. This means these systemic insecticides are present in pollen and nectar that pollinators can come in contact with when foraging. In addition, they have been found on neighboring flowers and grass (even at levels higher than the crops they were applied to), in nearby waterways, and they persist in the soil for long periods of time. The ability for these insecticides to escape into the environment and affect non-target organisms has sparked a lot of research interest into evaluating their implications and risks. These insecticides are emerging as being more toxic than other pesticides to bees. They also cause more lethal and sublethal effects on bumble bees compared to other pesticides. These make this group of insecticides the most studied class of insecticides for bees. source:link

3. The Niwot Ridge litter and woody debris dataset was collected from the National Ecological Observatory Network, which collectively includes 81 aquatic and terrestrial sites across 20 ecoclimatic domains. 32 of these sites sample forest litter and woody debris, and we will focus on the Niwot Ridge long-term ecological research (LTER) station in Colorado. Why might we be interested in studying litter and woody debris that falls to the ground in forests? Feel free to do a brief internet search if you feel you need more background information.

   Answer: The mission of Niwot Ridge long-term ecological research program as explained in its website is to better understand how complex mountain systems are changing and better predict the future of the many critical services these systems provide to all living downhill — in Boulder, in Colorado, and beyond. For this purpose, in order to have better understanding of the ecosystem, the NWT LTER program conducts study on the air, snow, water, soil, microbes, lakes, trees, flowers, and animals in the high mountains of the Colorado Rockies. The program measures, experiments and models how all these pieces fit together and have affected the health of our mountains over the last 40 years. In this context, specifically liter fall and fine woody debris data may be used to estimate annual Above-ground Net Primary Productivity and above-ground biomass at plot, site, and continential scales. They also provide essential data for understanding vegetative carbon fluxes over time source: Link.

4. How is litter and woody debris sampled as part of the NEON network? Read the NEON_Litterfall_UserGuide.pdf document to learn more. List three pieces of salient information about the sampling methods here:

Answer: Litter sample is collected from elevated traps and fine woody debris is collected from ground traps. The methodolgy of sample collection and the relevant information in relation to sampling is described below by referring the user guide. And they are categorized into three main points.

1.Sampling site/location selection: Litter and fine woody debris sampling is executed at terrestrial NEON sites that contain woody vegetation >2m tall. Sampling occurs only in tower plots.

- Locations of tower plots are selected randomly within the 90% flux footprint of the primary and secondary airsheds (and additional areas in close proximity to the airshed, as necessary to accommodate sufficient spacing between plots).
- In sites with forested tower airsheds, the litter sampling is targeted to take place in 20 40m x 40m plots.
- In sites with low-statured vegetation over the tower airsheds, litter sampling is targeted to take place in 4 40m x 40m tower plots (to accommodate co-located soil sampling) plus 26 20m x 20m plots.

2.Sampling technique:

- One litter trap pair (one elevated trap and one ground trap) is deployed for every 400 m2 plot area, resulting in 1-4 trap pairs per plot. In some cases, available space, plot spacing requirements, and/or the tower airshed size restricts the number of plots that can be sampled for litter below 20 (forested) or 30 (low-stature).
- Specifically, plot edges must be separated by a distance 150% of one edge of the plot (e.g., 40m x 40m Tower Base Plots must be 60m apart); plot centers must be greater than 50m from large paved roads and plot edges must be 10m from two track dirt roads; plot centers must be 50m from buildings and other non-NEON infrastructure; streams larger than 1m must not intersect plots.
- Trap placement within plots may be either targeted or randomized, depending on the vegetation. In sites with > 50% aerial cover of woody vegetation >2m in height, placement of litter traps is random and utlizes the randomized list of grid cell locations being utlized for herbaceous clip harvest and bryophyte sampling.
- In sites with < 50% cover of woody vegetation, sites with heterogeneously distributed, patchy, vegetation, trap placement is targeted such that only areas beneath qualifying vegetation are considered for trap placement.

3.Sampling time and frequency: Depending on the location of the traps, the frequency of sampling differs.

- Ground traps are sampled once per year.
- Target sampling frequency for elevated traps varies by vegetation present at the site, with frequent sampling (1x every 2weeks) in deciduous forest sites during senescence, and infrequent year-round sampling (1x every 1-2 months) at evergreen sites.
- At sites with deciduous vegetation or limited access during winter months, litter sampling of elevated traps may be discontinued for up to 6 months during the dormant season.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
dim(neonics.data)
```

```
## [1] 4623   30
```

- The Neonics data set has 4,623 observations and 30 variables.

6. Using the `summary` function on the "Effect" column, determine the most common effects that are studied. Why might these effects specifically be of interest?

```
summary(neonics.data$Effect)
```

```
##     Accumulation        Avoidance         Behavior     Biochemistry 
##               12              102              360               11 
##          Cell(s)      Development        Enzyme(s) Feeding behavior 
##                9              136               62              255 
##         Genetics           Growth        Histology       Hormone(s) 
##               82               38                5                1 
##    Immunological     Intoxication       Morphology        Mortality 
##               16               12               22             1493 
##       Physiology       Population     Reproduction 
##                7             1803              197 
```

Answer: The most common effects that are studied are Mortality and Population followed by behaviour, feeding behaviour, reproduction and development. Becuase these variables are the ones that can 1) incorporate the effects and consequences of most of the other varibles; 2) ultimately show the overall impact of the different concentration levels of the insecticides; and 3) serve as evidence by reinforcing each other.

7. Using the `summary` function, determine the six most commonly studied species in the dataset (common name). What do these species have in common, and why might they be of interest over other insects? Feel free to do a brief internet search for more information if needed.

```
summary(neonics.data$Species.Common.Name)
```

```
##                          Honey Bee                     Parasitic Wasp 
##                                667                                285 
##              Buff Tailed Bumblebee                Carniolan Honey Bee 
##                                183                                152 
##                         Bumble Bee                   Italian Honeybee 
##                                140                                113 
##                    Japanese Beetle                  Asian Lady Beetle 
##                                 94                                 76 
##                     Euonymus Scale                           Wireworm 
##                                 75                                 69 
##                  European Dark Bee                  Minute Pirate Bug 
##                                 66                                 62 
##               Asian Citrus Psyllid                      Parastic Wasp 
##                                 60                                 58 
##             Colorado Potato Beetle                    Parasitoid Wasp 
##                                 57                                 51 
##                Erythrina Gall Wasp                       Beetle Order
##                                 49                                 47
##        Snout Beetle Family, Weevil           Sevenspotted Lady Beetle
##                                 47                                 46
##                     True Bug Order              Buff-tailed Bumblebee
##                                 45                                 39
##                       Aphid Family                     Cabbage Looper
##                                 38                                 38
##               Sweetpotato Whitefly                      Braconid Wasp
##                                 37                                 33
##                       Cotton Aphid                     Predatory Mite
##                                 33                                 33
##             Ladybird Beetle Family                         Parasitoid
##                                 30                                 30
##                      Scarab Beetle                      Spring Tiphia
##                                 29                                 29
##                        Thrip Order               Ground Beetle Family
##                                 29                                 27
##                 Rove Beetle Family                      Tobacco Aphid
##                                 27                                 27
##                       Chalcid Wasp             Convergent Lady Beetle
##                                 25                                 25
##                      Stingless Bee                  Spider/Mite Class
##                                 25                                 24
##                Tobacco Flea Beetle                   Citrus Leafminer
##                                 24                                 23
##                    Ladybird Beetle                          Mason Bee
##                                 23                                 22
##                           Mosquito                      Argentine Ant
##                                 22                                 21
##                             Beetle         Flatheaded Appletree Borer
##                                 21                                 20
##               Horned Oak Gall Wasp                 Leaf Beetle Family
##                                 20                                 20
##                  Potato Leafhopper         Tooth-necked Fungus Beetle
##                                 20                                 20
##                       Codling Moth          Black-spotted Lady Beetle
##                                 19                                 18
##                       Calico Scale                Fairyfly Parasitoid
##                                 18                                 18
##                        Lady Beetle             Minute Parasitic Wasps
##                                 18                                 18
##                          Mirid Bug                   Mulberry Pyralid
##                                 18                                 18
##                           Silkworm                     Vedalia Beetle
##                                 18                                 18
##              Araneoid Spider Order                          Bee Order
##                                 17                                 17
##                     Egg Parasitoid                       Insect Class
##                                 17                                 17
##           Moth And Butterfly Order       Oystershell Scale Parasitoid
##                                 17                                 17
## Hemlock Woolly Adelgid Lady Beetle              Hemlock Wooly Adelgid
##                                 16                                 16
##                               Mite                        Onion Thrip
##                                 16                                 16
##              Western Flower Thrips                       Corn Earworm
##                                 15                                 14
##                  Green Peach Aphid                          House Fly
##                                 14                                 14
##                          Ox Beetle                 Red Scale Parasite
##                                 14                                 14
##                 Spined Soldier Bug              Armoured Scale Family
##                                 14                                 13
##                   Diamondback Moth                      Eulophid Wasp
##                                 13                                 13
##                  Monarch Butterfly                      Predatory Bug
##                                 13                                 13
##              Yellow Fever Mosquito                Braconid Parasitoid
##                                 13                                 12
##                       Common Thrip       Eastern Subterranean Termite
##                                 12                                 12
##                             Jassid                         Mite Order
##                                 12                                 12
##                          Pea Aphid                   Pond Wolf Spider
##                                 12                                 12
##           Spotless Ladybird Beetle             Glasshouse Potato Wasp
##                                 11                                 10
##                           Lacewing            Southern House Mosquito
##                                 10                                 10
##            Two Spotted Lady Beetle                         Ant Family
##                                 10                                  9
##                       Apple Maggot                            (Other)
##                                  9                                670
```

Answer: The six most common studied species in the data set by their common name are Honey Bee (667),Parasitic Wasp (285), Buff Tailed Bumblebee (183),Carniolan Honey Bee (152),Bumble Bee (140) and Italian Honeybee (113). These insects are beneficial to human beings and the environment by playing a significant role in the process of **pollination** and, pushing/fighting and killing other harmful insects like aphids. For their significant and beneficial contribution for the healthy of the ecosystem, they have become the subject of interest over other insects.

8. Concentrations are always a numeric value. What is the class of Conc.1..Author. in the dataset, and why is it not numeric?

```
class(neonics.data$Conc.1..Author.)
```

```
## [1] "factor"
```

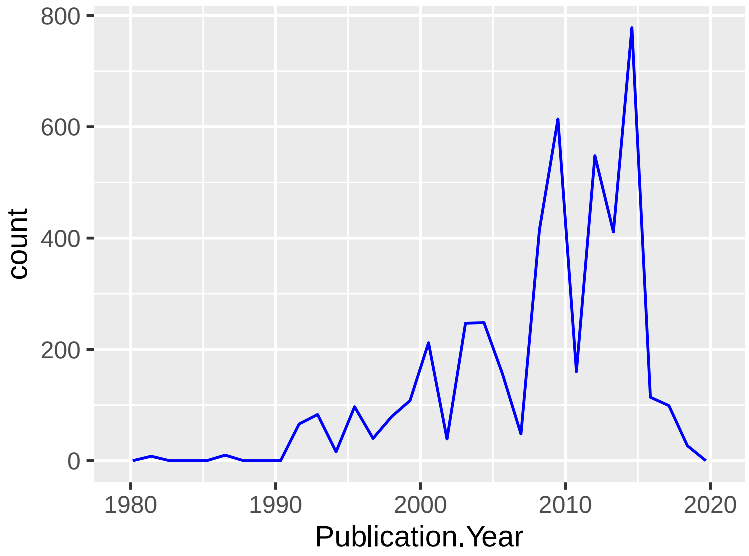
Answer: The class of Conc.1..Author in the dataset is "factor". As mentioned in the user guide, the code CONC TYPE is used in the publication to report values that the author refers to the concentration as active ingredient, active principle, active substance (A.S.), acid equivalent or various grades of reagents (ie., Analytical, Reagent or Technical). Thus, a value in the publication may be reported as "AI kg/ha", "AE kg/ha" or "kg AI/ha" which give the variable a categorical nature to be more meaningful rather than numeric. Why it is not numeric?

Explore your data graphically (Neonics)

9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```
ggplot(neonics.data) +
 geom_freqpoly(aes(x = Publication.Year), color = "blue")
```

```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```

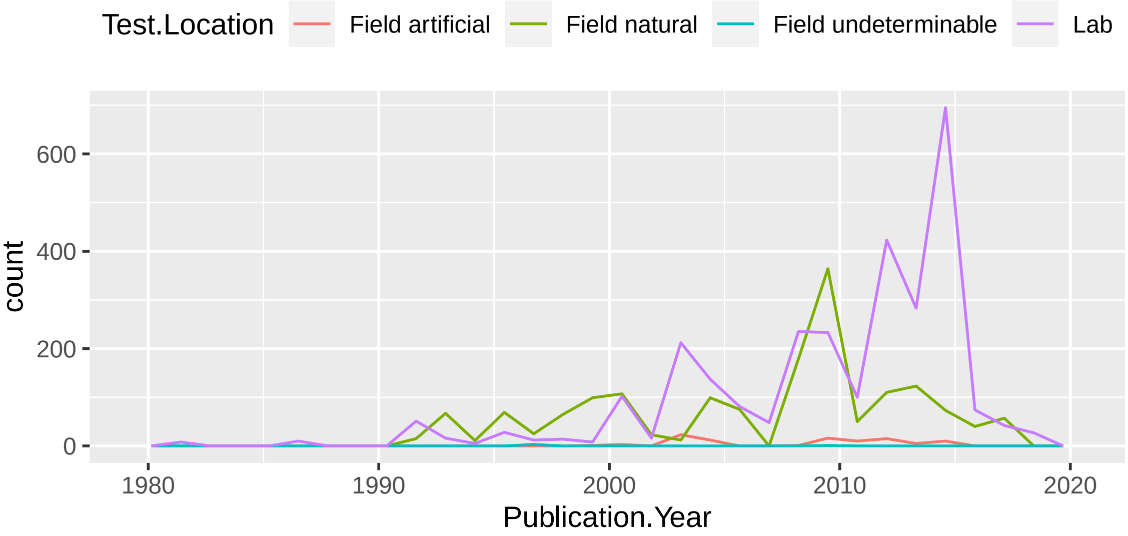


10. Reproduce the same graph but now add a color aesthetic so that different Test.Location are displayed as different colors.

```
ggplot(neonics.data) +
 geom_freqpoly(aes(x = Publication.Year, color =  Test.Location, bin = 50)) +
  theme(legend.position = "top")
```

```
## Warning: Ignoring unknown aesthetics: bin
```

```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```

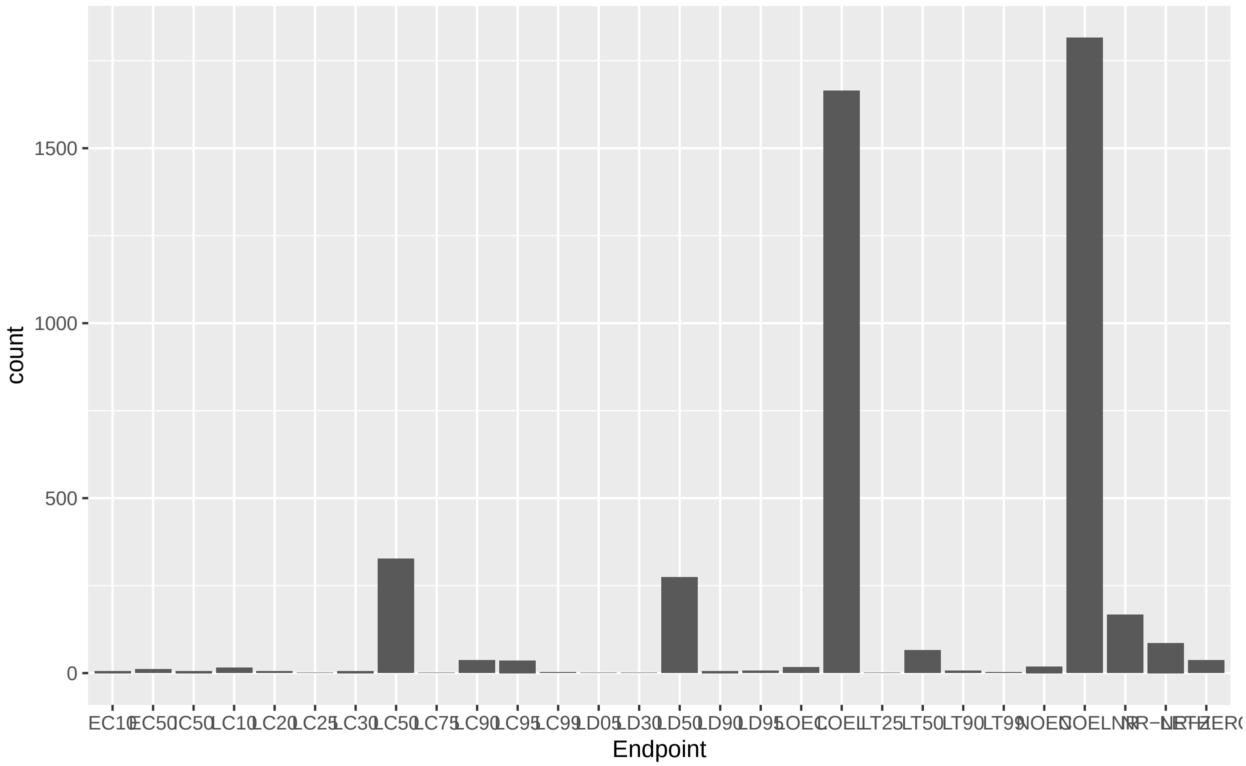


Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The test locations are categorized in to four groups: Field artificial, Field natural, Field undeterminable and Laboratory. Of these groups, the Field artificial test location come to use around the end of 1990s, reached its small peak at the begining of 2000 and showed a declined trend until 2015/16 after which they are not used. The Field undeterminable category remaind constantly at the level of zero througout the publication periods used. The most common test locations are field natural and laboratory, the laboratory showing a dramatic increase overtime followed by the field natural. These test locations differ overtime in that the Laboratory is the dominant test location in recent times. However, both showed a similar trend of an increase and decrease with alternate dominance of one over the other at differen times (at the begining of 2000 and 2010, and until near to 2020 significant dominance by the laboratory test location; and in the 1990s to 2000 and at the end of 2010 dominance by the Field natural test location).

11. Create a bar graph of Endpoint counts. What are the two most common end points, and how are they defined? Consult the ECOTOX_CodeAppendix for more information.

```
ggplot(neonics.data, aes(x = Endpoint)) +
  geom_bar()
```



```
summary(neonics.data$Endpoint) # To clearly see the Endpoint labeling and their frequency
```

```
##    EC10    EC50    IC50    LC10    LC20    LC25    LC30    LC50    LC75    LC90
##       6      11       6      15       5       1       6     327       1      37
##    LC95    LC99    LD05    LD30    LD50    LD90    LD95    LOEC    LOEL    LT25
##      36       2       1       1     274       6       7      17    1664       1
##    LT50    LT90    LT99    NOEC    NOEL      NR NR-LETH NR-ZERO
##      65       7       2      19    1816     167      86      37
```

Answer: As shown in the bar graph and preciesly indicated using the summary function, the two most common end points are LOEL (1664) and NOEL (1816). As indicated in the ECOTOX_CodeAppendix: * LOEL stands for Lowest-observable-effect-level which is lowest does (concentration) producing effects that were significantly different from responses of controls (LOEAL - Lowest-observed-adverse-effect-level /LOEC - Lowest-observed-effect-concentration) (as reported by authors) * NOEL stands for No-observable-effect-level which means highest dose (concentration) producing effects not significantly different from responses of controls according to author's reported statistical test (NOEAL - No-observed-adverse-effect-level/NOEC - No-observed-effect-concentration)

Explore your data (Litter)

12. Determine the class of collectDate. Is it a date? If not, change to a date and confirm the new class of the variable. Using the `unique` function, determine which dates litter was sampled in August 2018.

```
class(litter.data$collectDate)
```

```
## [1] "factor"
```

- The class of collectDate is “factor”.

```
head(litter.data$collectDate)
```

```
## [1] 2018-08-02 2018-08-02 2018-08-02 2018-08-02 2018-08-02 2018-08-02
## Levels: 2018-08-02 2018-08-30
```

```
litter.data$collectDate <- as.Date(litter.data$collectDate, format = "%Y-%m-%d")
```

- The formate of the date used in the data set is yyyy-mm-dd.

```
class(litter.data$collectDate)
```

```
## [1] "Date"
```

- Now the class of the variable (collectDate) become “Date”

```
unique(litter.data$collectDate)
```

```
## [1] "2018-08-02" "2018-08-30"
```

- The dates litter was sampled in August 2018 were, the 2nd and 30th day of the month. Or The litter sample collection dates were August 02 and August 30, 2018.

13. Using the `unique` function, determine how many plots were sampled at Niwot Ridge. How is the information obtained from `unique` different from that obtained from `summary`?

```
unique(litter.data$plotID)
```

```
##  [1] NIWO_061 NIWO_064 NIWO_067 NIWO_040 NIWO_041 NIWO_063 NIWO_047 NIWO_051
##  [9] NIWO_058 NIWO_046 NIWO_062 NIWO_057
## 12 Levels: NIWO_040 NIWO_041 NIWO_046 NIWO_047 NIWO_051 NIWO_057 ... NIWO_067
```

- The unique plot Ids in the data set used for sampling are 12. Thus,the number of plots sampled at Niwot Ridge is 12.

```
summary(litter.data$plotID)
```

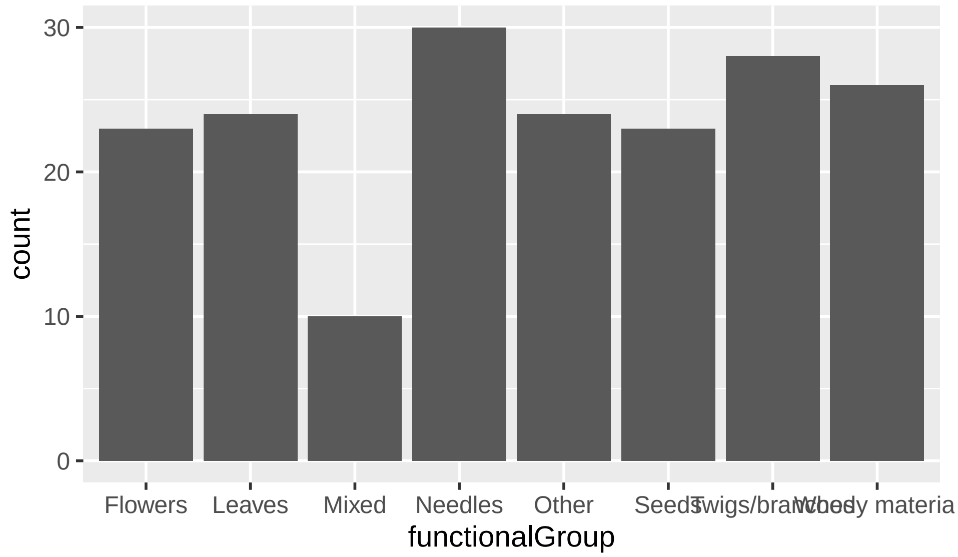
```
## NIWO_040 NIWO_041 NIWO_046 NIWO_047 NIWO_051 NIWO_057 NIWO_058 NIWO_061
##       20       19       18       15       14        8       16       17
## NIWO_062 NIWO_063 NIWO_064 NIWO_067
##       14       14       16       17
```

- The total number of plots used for the sample including the frequency of the plots used for sampling.

  Answer: The number of plots sampled at Niwot Ridge is 12. The samples taken repeatdly from those 12 plots in a different date. The information obtained using the ‘unique’ function is different from the ‘summary’ function in such a way that the ‘unique’ function gives us unique plot Id by removing duplicated plot Ids, whereas the ‘summary’ function gives us the total plots keeping and counting the duplicated plots where samples are taken from repeatdly.

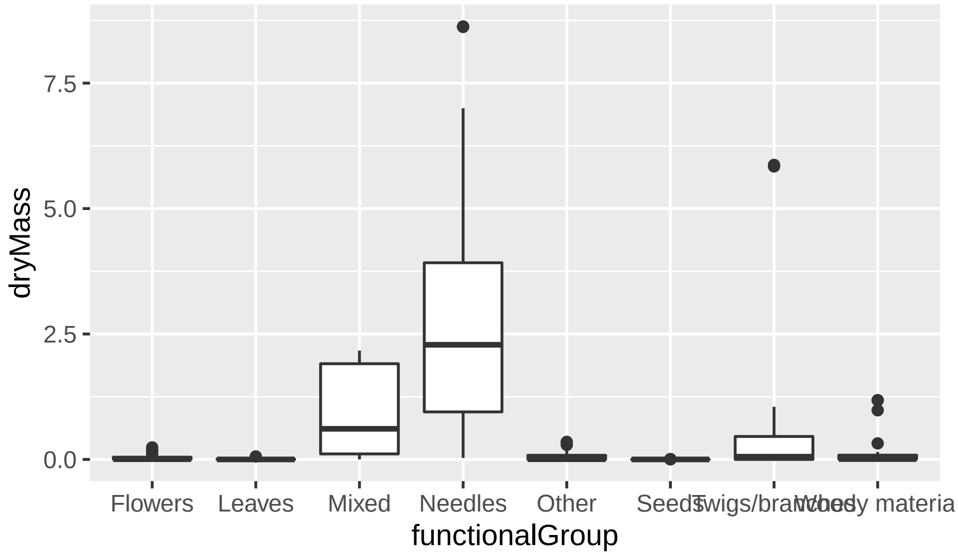
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
ggplot(litter.data, aes(x = functionalGroup)) + geom_bar()
```



15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
ggplot(litter.data, aes(x =   functionalGroup, y = dryMass)) + geom_boxplot()
```



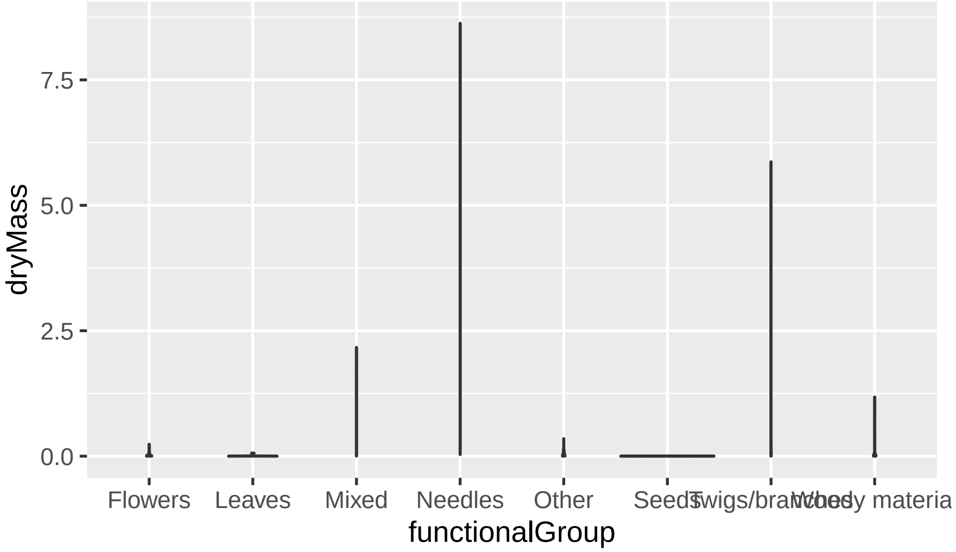
```
ggplot(litter.data) +
geom_violin(aes(x = functionalGroup, y = dryMass), draw_quantiles = c(0.25, 0.5, 0.75))
```

```
## Warning in regularize.values(x, y, ties, missing(ties), na.rm = na.rm):
```

```
## collapsing to unique 'x' values

## Warning in regularize.values(x, y, ties, missing(ties), na.rm = na.rm):
## collapsing to unique 'x' values

## Warning in regularize.values(x, y, ties, missing(ties), na.rm = na.rm):
## collapsing to unique 'x' values
```



Why is the boxplot a more effective visualization option than the violin plot in this case?

> Answer: The boxplot clearly indicates the distribution of the observations (drymass) in each functional group. It does this by showing the median, the IQR (the first and third quartiles), the maximum, the minimum and the outlier observations. The violin plot in this case on the other hand doesn't show these values in a meaningful manner, meaning it doesn't show the range of values and the distribution or shape of the data within each functional group as expected.

What type(s) of litter tend to have the highest biomass at these sites?

> Answer: Needles have the highest drymass or biomass at these sites followed by mixed functional group and twigs/branches respectively.